Classification with Valid and Adaptive Coverage

徐少东

中国人民大学统计学院

2021 年 12 月 1 日

- 假设有 n 个样本 $\{(X_i, Y_i)\}_{i=1}^n$，其中 $X_i \in \mathbb{R}^p$ 为样本的特征，$Y_i \in \mathcal{Y} = \{1, 2, ..., C\}$ 为样本的标签；
- 样本满足可交换假设，样本所服从的未知分布用 P_{XY} 表示；
- 在特定覆盖水平 $1-\alpha$ 下，构建新样本 (X_{n+1}, Y_{n+1}) 标签的预测集合 $\hat{C}_{n,a} \subseteq \mathcal{Y}$，其中 X_{n+1} 已知而 Y_{n+1} 未知。
- 预测集合满足边际覆盖（marginal coverage）

$$\mathbb{P}\left[Y_{n+1} \in \hat{C}_{n,a}(X_{n+1})\right] \geq 1-\alpha.$$

- 边际覆盖在实际应用中可以很好地实现，但边际覆盖并不能推出条件覆盖：

$$\mathbb{P}\left[Y_{n+1} \in \hat{C}_{n,a}(x) | X_{n+1} = x\right] \geq 1 - \alpha.$$

- 条件覆盖希望对于特定的 X 的观察值实现有效覆盖。
- 没有较强的模型假设，条件覆盖理论上无法实现（Barber et al., 2019）。

边际覆盖和条件覆盖的区别在哪里？

- 假定每一个样本 i 对应一个病人，X_i 代表这个病人的相关协变量（年龄、家族病史等），Y_i 表示一个定量的结果（比如，服用某种药物后血压会降低多少）；
- 当一个新的病人进来，他的相关变量为 X_{n+1}，医生想预测他的结果 Y_{n+1}；
- 在边际覆盖意义下，医生的预测对于接下来所有可能到来的病人的平均值，有 95% 的机会是正确的；

边际覆盖和条件覆盖的区别在哪里？

- 假如有 95% 以上的患者都是 25 岁以上，那么如果新来的患者是 25 岁以下的，医生在边际覆盖意义下作出的 95% 预测区间可能对这个患者来说，正确率是 0%。
- 因此条件覆盖的定义更强，他要求不管进来一个什么样的样本，都得保证作出的预测对于这个样本来说，正确率不小于 95%。

1. 提出一种分类方法在边际覆盖下能够理论证明其有效性；
2. 同时可以估计该方法的条件覆盖率；
3. 并且保证方法的准确性，体现在 $\hat{C}$ 尽可能的小。

- 在边际覆盖下提出的有关分类的共形推断方法有很多 (Hechtlinger et al., 2018; Sadinle et al., 2019; Vovk et al., 2005)。
- 最初为 Vladimir Vovk 提出，主要学者为 Larry Wasserman、Jing Lei 等。

Oracle 分类器

假设我们已知条件分布 $P_{Y|X}$ 来对 oracle 性质进行探究。

- 定义最优的预测集合 $C_\alpha^{oracle}(X_{n+1})$;
- 对于任意的 $x \in \mathbb{R}^p$，令 $\pi_y(x) = \mathbb{P}[Y = y|X = x]$，即我们可以得到给定 $X = x$ 下的任意一个标签的条件概率;
- 定义 $\pi_y(x)$ 的次序统计量 $\pi_{(1)}(x) \geq \pi_{(2)}(x) \geq \cdots \geq \pi_{(C)}(x)$。
- 对于任意 $\tau \in [0, 1]$，定义广义条件分位函数:

$$L(x; \pi, \tau) = min\{c \in \{1, ..., C\} : \pi_{(1)}(x) + \pi_{(2)}(x) + \cdots + \pi_{(C)}(x) \geq \tau\}$$

Oracle 分类器

预测集合为

$$C_\alpha^{oracle+}(x) = \{\text{前} L(x;\pi,1-\alpha)\text{个最大的}\pi_y(x)\text{对应的}y\text{的 indexes}\}$$

即在 $1-\alpha$ 水平下给定 $X = x$ 包含响应变量最少的集合。

Oracle 分类器

举例来说，如果 $\pi_1(x) = 0.3$，$\pi_2(x) = 0.6$ 以及 $\pi_3(x) = 0.1$，则 $L(x; 0.9) = 2$，取最大的两个，得到 $C_{0.1}^{oracle}(x) = \{1, 2\}$。而 $L(x; 0.5) = 1$，即取最大的一个，得到 $C_{0.5}^{oracle}(x) = \{2\}$。

Oracle 分类器

定义一个新的函数 $\mathcal{S}$ 来计算预测集合，加入一个新的残暑 $u \in [0,1]$。一般情况下，

$$\mathcal{S}(x, u; \pi, \tau) = C_{1-\tau}^{oracle}(x)$$

但是当 $u \leq V(x; \pi, \tau)$ 时，

$$\mathcal{S}(x, u; \pi, \tau) = C_{1-\tau}^{oracle}(x)/\text{最大的}\pi_y(x)\text{对应的}y$$

其中

$$V(x; \pi, \tau) = \frac{1}{\pi_{(L(x;\pi,\tau))}(x)} \left[\sum_{c=1}^{L(x;\pi,\tau)} \pi_{(c)}(x) - \tau \right].$$

Oracle 分类器

令 u 服从一个均匀分布，可以得到一个更紧的随机预测集合：

$$C_\alpha^{oracle}(x) = \mathcal{S}(x, U; \pi, 1-\alpha).$$

其中 $U \sim Uniform(0,1)$ 并且与其他变量独立。
这样得到的集合是条件覆盖水平在 $1-\alpha$ 下最小的随机预测集合。

Oracle 分类器

在上述例子中，$L(x.0.5) = 1$，有 (0.6-0.5)/0.6=1/6 的概率 $C_{0.5}^{oracle}(x) = \emptyset$，有 5/6 的概率 $C_{0.5}^{oracle}(x) = \{2\}$。

- 这篇文章使用训练好的分类器来估计未知的条件分布 $P_{Y|X}$;
- 所提出的方法的主要优势在于可以与任何黑箱预测模型进行搭配，比如神经网络等;
- 唯一的限制是分类器需要可交换的处理所有的样本;
- 许多现成的估计 $\pi_y(x)$ 的方法我们可以直接拿来使用，借助基于 oracle 分析得到的算法，来获得预测集合，并保证覆盖率。

- 使用 $\hat{\pi}_y(x)$ 来替代 $\pi_y(x)$ 不能保证前面所推出的覆盖率能达到，因为这个估计可能不够准确；
- 即 $\tau = 1 - \alpha$ 不能保证条件覆盖率能达到 $-\alpha$；
- 但是我们可以在 hold-out 数据中计算覆盖率，然后反过来调整 τ；
- 然后选择能够让覆盖率达到 $1 - \alpha$ 的最小的 τ。

基本思想

假设我们有一个分类器，并得到了估计 $\hat{\pi}_y(x)$。这里只假定 $\hat{\pi}_y(x)$ 是标准化的。
然后我们将 π 放入上文由 oracle 条件推出的预测集合计算方法中，与 oracle 情况下不同的是，阈值 τ 需要使用独立于训练集的 hold-out 数据精心调整。

基本思想

定义函数 E，输入 $x, y, u, \hat{\pi}$，这个函数输出使得 $\mathcal{S}(x, u; \pi, \tau)$ 在给定 $X = x$ 下包含标签 y 的 τ 的最小值。
这就是我们所需的 generalized inverse quantile conformity score function：

$$E(x, y, u; \hat{\pi}) = \min\{\tau \in [0, 1] : y \in \mathcal{S}(x, u; \pi, \tau)\}.$$

有了这个函数，我们就可以在 hold-out 样本 (X_i, Y_i) 上计算共形度得分，记作 $E_i = E(X_i, Y_i, U_i; \hat{\pi})$。在理想情况下，即 $\hat{\pi} = \pi$ 时，以 X 为条件，E_i 是均匀分布的。

基本思想

$$E(x, y, u; \hat{\pi}) = \min\{\tau \in [0, 1] : y \in \mathcal{S}(x, u; \pi, \tau)\}.$$

可以将其视作一种特殊的 p 值。
这个特性使得本文得到的得分可在不同样本中有可比性。
之前所提出的共形度得分即使是在 oracle 情况下，往往在 X 取值不同的时候有不同的分布。而本文方法是在 $\hat{\pi}$ 下即可实现。
使用 $\{E_i\}_{i\in\mathcal{I}_2}$ 的 $1-$ 分位点就可以得到 τ 应该的取值，其中 $\mathcal{I}_2$ 是用来调整 τ 的 hold-out 数据，这部分数据没有用于估计 π。

算法

Algorithm 1: Adaptive classification with split-conformal calibration

Input: data $\{(X_i, Y_i)\}_{i=1}^n$, X_{n+1}, black-box learning algorithm $\mathcal{B}$, level $\alpha \in (0, 1)$.
Randomly split the training data into 2 subsets, $\mathcal{I}_1, \mathcal{I}_2$.
Sample $U_i \sim \text{Uniform}(0, 1)$ for each $i \in \{1, \ldots, n+1\}$, independently of everything else.
Train $\mathcal{B}$ on all samples in $\mathcal{I}_1$: $\hat{\pi} \leftarrow \mathcal{B}(\{(X_i, Y_i)\}_{i \in \mathcal{I}_1})$.
Compute $E_i = E(X_i, Y_i, U_i; \hat{\pi})$ for each $i \in \mathcal{I}_2$, with the function E defined in (7).
Compute $\hat{Q}_{1-\alpha}(\{E_i\}_{i \in \mathcal{I}_2})$ as the $\lceil (1-\alpha)(1+|\mathcal{I}_2|) \rceil$th largest value in $\{E_i\}_{i \in \mathcal{I}_2}$.
Use the function $\mathcal{S}$ defined in (5) to construct the prediction set at X_{n+1} as:

$$\hat{\mathcal{C}}^{\text{SC}}_{n,\alpha}(X_{n+1}) = \mathcal{S}(X_{n+1}, U_{n+1}; \hat{\pi}, \hat{Q}_{1-\alpha}(\{E_i\}_{i \in \mathcal{I}_2})). \tag{8}$$

Output: A prediction set $\hat{\mathcal{C}}^{\text{SC}}_{n,\alpha}(X_{n+1})$ for the unobserved label Y_{n+1}.

算法性质

如果样本 (X_i, Y_i)，对于 $i \in \{1, ..., n+1\}$ 都是可交换的，并且分类器 $\mathcal{B}$ 不受输入样本的顺序影响，算法 1 的输出满足边际覆盖：

$$\mathbb{P}\left[Y_{n+1} \in C^{SC}_{n,\alpha}(X_{n+1})\right] \geq 1 - \alpha.$$

并且，如果得分 E_i 是几乎处处唯一的，边际覆盖接近是紧的：

$$\mathbb{P}\left[Y_{n+1} \in C^{SC}_{n,\alpha}(X_{n+1})\right] \leq 1 - \alpha + \frac{1}{|\mathcal{I}_2| + 1}.$$

.

算法

Algorithm 2: Adaptive classification with CV+ calibration

Input: data $\{(X_i, Y_i)\}_{i=1}^n$, X_{n+1}, black-box $\mathcal{B}$, number of splits $K \leq n$, level $\alpha \in (0,1)$.
Randomly split the training data into K disjoint subsets, $\mathcal{I}_1, \ldots, \mathcal{I}_K$, each of size n/K.
Sample $U_i \sim \text{Uniform}(0,1)$ for each $i \in \{1, \ldots, n+1\}$, independently of everything else.
for $k \in \{1, \ldots, K\}$ **do**
    Train $\mathcal{B}$ on all samples except those in $\mathcal{I}_k$: $\hat{\pi}^k \leftarrow \mathcal{B}(\{(X_i, Y_i)\}_{i \in \{1,\ldots,n\} \setminus \mathcal{I}_k})$.
end
Use the function E defined in (7) to construct the prediction set $\hat{\mathcal{C}}_{n,\alpha}^{\text{CV+}}(X_{n+1})$ as:

$$\hat{\mathcal{C}}_{n,\alpha}^{\text{CV+}}(X_{n+1}) = \Bigg\{ y \in \mathcal{Y} : \sum_{i=1}^{n} \mathbf{1}\left[E(X_i, Y_i, U_i; \hat{\pi}^{k(i)}) < E(X_{n+1}, y, U_{n+1}; \hat{\pi}^{k(i)}) \right] < (1-\alpha)(n+1) \Bigg\}, \tag{11}$$

where $k(i) \in \{1, \ldots, K\}$ is the fold containing the ith sample.
Output: A prediction set $\hat{\mathcal{C}}_{n,\alpha}^{\text{CV+}}(X_{n+1})$ for the unobserved label Y_{n+1}.

homogeneous conformal classification

之前提出的方法构建的预测区间大部分基于一个简单的规则：

$$\hat{C}(x;t) = \{y \in \mathcal{Y} : \hat{f}(y|x) \geq t\},$$

但由于所有的样本都使用同一个阈值 t，所以条件覆盖率会低于预定水平。

quantile regression

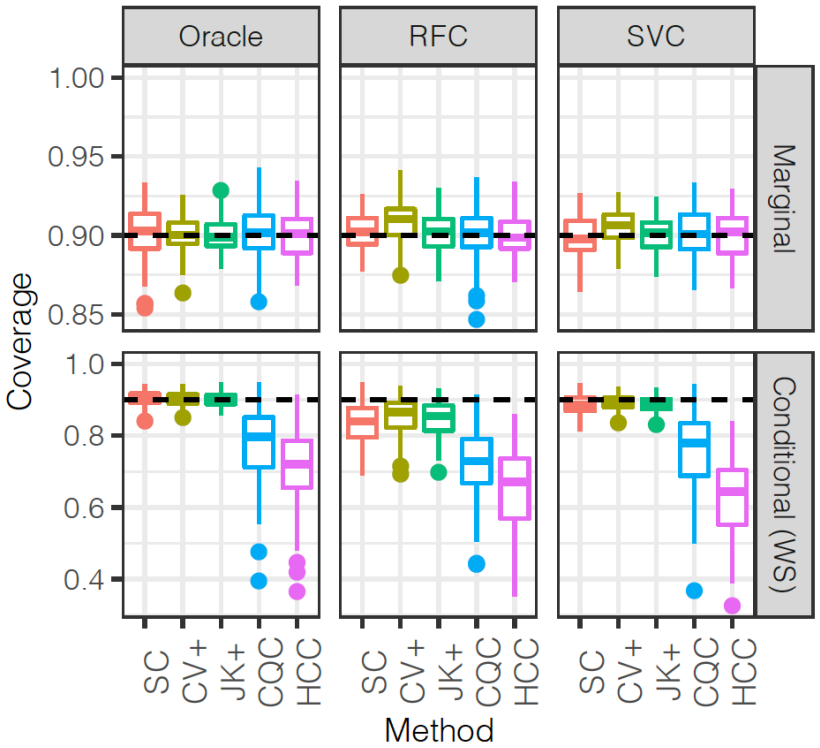
这种方法有两个缺陷：

1. 首先，它涉及额外的数据拆分以避免过度拟合，这往往会降低模型的效果；
2. 其次，它的理论渐近最优性比我们的弱，因为它需要两个黑箱的一致性而不是一个。

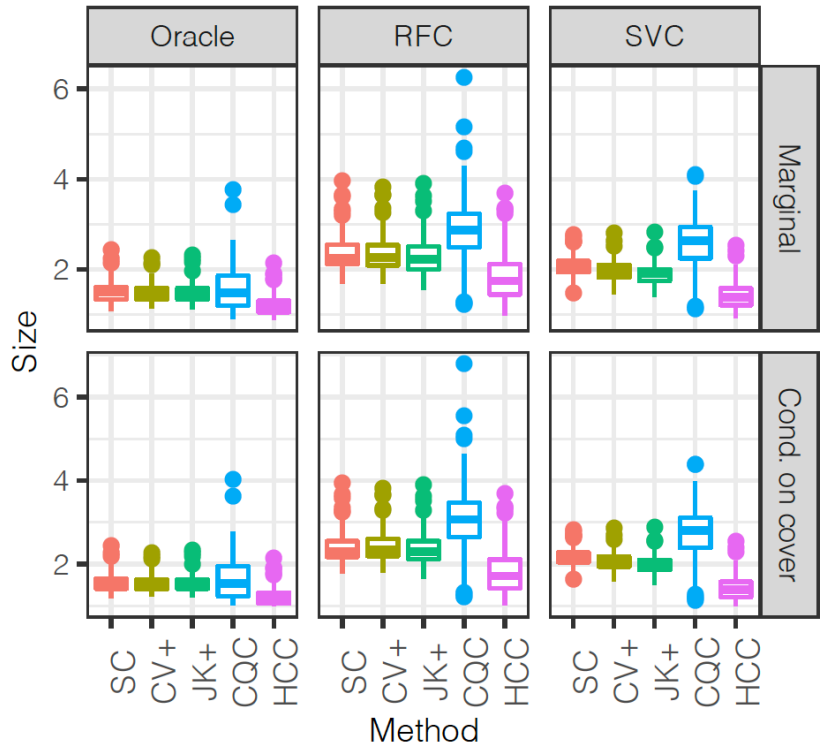
模拟设置

- 数据 $X \in \mathbb{R}^p$，其中 $p = 10$；
- X_1 是信号变量，X_1 以 0.2 概率取 1，以 0.8 的概率取-8；
- 其他变量为噪音变量，都独立地服从标准正态分布；
- 标签 $Y \in \{1, ..., 10\}$ 在给定 $X = x$ 时的条件分布为一个 multinomial 分布，其中权重 $w_j(x)$ 定义为 $w_j(x) = z_j(x) / \sum_{j'=1}^{p} z_{j'}(x)$，其中 $z_j(x) = exp(x^T \beta_j)$，每一个 $\beta_j \in \mathbb{R}^p$ 是从独立的标准正态分布中抽样得到。

模拟结果


Oracle
RFC
SVC
Marginal
Conditional (WS)
Coverage
1.00
0.95
0.90
0.85
1.0
0.8
0.6
0.4
SC
CV+
JK+
CQC
HCC
Method

(a)


Oracle
RFC
SVC
Marginal
Cond. on cover
Size
6
4
2
SC
CV+
JK+
CQC
HCC
Method

(b)